Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Emergence of Online Community Leadership

Steven L. Johnson, Hani Safadi, Samer Faraj

To cite this article:
Steven L. Johnson, Hani Safadi, Samer Faraj (2015) The Emergence of Online Community Leadership. Information Systems Research 26(1):165-187. https://doi.org/10.1287/isre.2014.0562

Information Systems Research
Vol. 26, No. 1, March 2015, pp. 165–187
ISSN 1047-7047 (print) | ISSN 1526-5536 (online)

http://dx.doi.org/10.1287/isre.2014.0562

The Emergence of Online Community Leadership

Steven L. Johnson
Fox School of Business, Temple University, Philadelphia, Pennsylvania 19122,
steven@temple.edu

Hani Safadi
Howe School of Technology Management, Stevens Institute of Technology, Hoboken, New Jersey 07030,
hanisaf@gmail.com

Samer Faraj
Desautels Faculty of Management, McGill University, Montréal, Québec H3A 1G5, Canada,
samer.faraj@mcgill.ca

Compared to traditional organizations, online community leadership processes and how leaders emerge are not well studied. Previous studies of online leadership have often identified leaders as those who administer forums or have high network centrality scores. Although communication in online communities occurs almost exclusively through written words, little research has addressed how the comparative use of language shapes community dynamics. Using participant surveys to identify leading online community members, this study analyzes a year of communication network history and message content to assess whether language use differentiates leaders from other core community participants. We contribute a novel use of textual analysis to develop a model of language use to evaluate the utterances of all participants in the community. We find that beyond communication network position—in terms of formal role, centrality, membership in the core, and boundary spanning—those viewed as leaders by other participants, post a large number of positive, concise posts with simple language familiar to other participants. This research provides a model to study online language use and points to the emergent and shared nature of online community leadership.

Keywords: online communities; leadership; natural language processing; knowledge management; network analysis; computer-mediated communication and collaboration

History: Natalia Levina, Senior Editor; Jonathon Cummings, Associate Editor. This paper was received on June 25, 2012, and was with the authors 15 months for 3 revisions.

> The key to successful leadership is influence, not authority.
> —Kenneth H. Blanchard

Introduction

Supported by the widespread use of social media, online communities have rapidly emerged as essential new forms of organizing (Benkler 2006, Kraut and Resnick 2011, Preece 2000). Online communities are large collectivities, where members with shared goals and interests interact primarily via the Internet (Sproull and Arriaga 2007). They bring together thousands of strangers across national, geographic, time zone, and organizational boundaries. Some communities focus on sustaining social ties and friendship (e.g., Facebook). Others serve as platforms for knowledge integration (e.g., Wikipedia), for sharing creative output (e.g., YouTube), for open source software development (e.g., Linux), or for answering questions (e.g., Quora.com). There is literally an online community to support every kind of interest, self-identified group, or creative endeavor.

In spite of the rapid growth of this new organizational form, research has been slow to examine the points of commonality and difference between traditional organizations and online communities. Principally, little is known about the rich diversity of forms of online collaboration, how they are structured, and how they sustain themselves (Faraj et al. 2011). Many of these communities are characterized by a core/periphery structure suggestive of interactions typical of communities of practice (Collier and Kraut 2012, Wasko et al. 2009). Members' decisions to participate may be due to a variety of intrinsic and extrinsic motivations (Kankanhalli et al. 2005, Lakhani and von Hippel 2003). Their level of engagement is affected by the strength of their identification with the group and the kind of interpersonal bonds they develop (Ren et al. 2012). Furthermore, in production- or expertise-based communities, continued participation is linked to the depth of embeddedness in the social practice encompassing the communal activity (von Krogh et al. 2012a, Wasko and Faraj 2005). Online communities are often characterized by high turnover, fluid boundaries, expertise-based authority, and emergent roles (Faraj et al. 2011, Ren et al. 2007).

In this paper, we focus on leadership processes in online communities. Whereas thousands of published works have enriched the understanding of organizational leadership, much less is known as to what constitutes effective leadership online. Given the lack of face-to-face communication, the mediated nature of interactions, and the primacy of text-based asynchronous exchanges, online leadership is bound to differ in some substantial ways from more familiar in-person and synchronous settings. Early findings indicate that leadership roles are more informal and emergent (Butler et al. 2007, Collier and Kraut 2012). Network position at the center of the exchanges seems to matter greatly (Sutanto et al. 2011). Leaders are heavily involved in the social practice of the community, its core mission, and its core activity whether it is developing code in open source software or answering questions in an expertise-based community (Dahlander and Frederiksen 2012, von Krogh et al. 2012a, Wasko and Faraj 2005). What makes someone a leader online, given the relative weakness of hierarchy and bottom-up governance structure, remains an open research question (O'Mahony and Ferraro 2007, von Krogh et al. 2012b, Yoo and Alavi 2004).

Because there is no one true definition for leadership, definitions should be made consistent with a study's substantive and methodological approach (Bass and Bass 2008). We define an online community leader as a participant recognized by other participants as influential in what the community does or how it does it (Yukl 2010). As such, online community leadership is not a stable global designation. Formally occupying a defined role of authority is neither a necessary nor sufficient condition for demonstrating online community leadership. Although interactions in the form of message posts are visible to all participants, different participants read different content and interpret content differently. Online communities are characterized by fluid structures and shifting membership, and are sustained through the voluntary contribution of members. Their structure is dependent on active posting and is therefore constituted by the interactions. Thus, a premise of this paper is that online community leadership is a local designation both distinct from formal roles and emerging from observable interactions.

We are motivated by the goal of understanding what leaders actually do in online communities. We consider both the network position resulting from a leader's interactions as well as the characteristics of a leader's written communications. Therefore, we first examine how communication network position–in terms of formal role, centrality, membership in the core, and boundary spanning–affects the likelihood of being seen as a leader. Then we contribute a novel use of textual analysis to develop a language model of utterances in the community to evaluate how convergent or divergent leader language is compared to that of the community as a whole. Our findings suggest that the most influential participants of any online community, those viewed as leaders by other participants, are not just among the most central, but also post a large number of positive, concise posts with simple language familiar to other participants.

Three important innovations strengthen our results. First, the leaders in our study are identified by community members rather than deduced based on structural position or behavior, as has been common practice in the majority of online leadership studies. Second, we contribute methodologically by comparing the identified leaders to a comparable set of participants that post an equivalent number of messages rather than attempting to compare to an average participant of the community—something futile given that an "average" participant is nonrepresentative in online communities characterized by a power law distribution of participation (Faraj and Johnson 2011, Newman 2003). Finally, our language model offers a sophisticated set of semantic and syntactic tools for the analysis of community discourse, again an advance over previous research models based on frequencies of often preidentified words.

Conceptualizations of Leadership in Online Communities

We argue that a synthesis of organizational leadership theories is required for a deeper understanding of leadership processes in online communities. In drawing on theories of leadership that emphasize behaviors associated with leadership, we identify four as particularly relevant to leadership in online communities: functional leadership, leader–member exchange (LMX), shared leadership, and communication as constitutive of organizing (CCO). Next, we discuss each theory and how it can be applied to online settings.

Functional leadership theory identifies behaviors that distinguish successful leaders and looks for associations between effective leadership and the functions performed (Burke et al. 2006). Leadership is not considered a personal characteristic, but rather can be identified as a set of behaviors that contribute to a group's goals and operation. The theory focuses on general leader behaviors and elaborates how they influence team processes and outcomes (Hackman and Walton 1986, Morgeson et al. 2010). Like functional leadership theory, we are also interested in what leaders do in online communities. Nonetheless, two specific limitations require the adaptation of leadership theory to the online communities. First, given that online communities lack the stable structure and visible leadership that characterize traditional organizations and teams, it is not clear that a focal leader can be identified a priori (Butler et al. 2007). Second, both the functions of leadership (Butler et al. 2007) and outcomes of successful leadership (Huffaker 2010, Zhu et al. 2012) are different in online communities.

Leader–member exchange theory focuses on the dyadic relationship between the leader and the team member. LMX theory assumes that the characteristics of the interactions between a leader and each of their team members is correlated with leadership processes and organizational outcomes (Gerstner and Day 1997). Given that team members have diverse abilities, commitments, roles, and responsibilities, a leader can improve team function by engaging in customized interactions with each team member based on their potential contribution to the team task (Graen and Uhl-Bien 1995). We share this view that online leadership behaviors are contingent and situated. Yet, the ability to directly apply LMX theory is constrained by large differences in group size between task-oriented workgroups and online communities. Specifically, given that online communities typically contain thousands of members, direct leadership relationships are bound to be tenuous and lacking in direct influence when compared with smaller, organizationally embedded teams in traditional face-to-face settings (Kiesler et al. 2011). Indeed, communication in online communities tends to include not only patterns of direct dyadic reciprocation but also generalized exchange patterns of indirect reciprocation (Faraj and Johnson 2011). Furthermore, because communication in online communities is typically open—all participants can read all communication—the ability for a leader to engage in differentiated direct exchange is diminished.

Shared leadership theory emphasizes the need for members to colead each other. Also known by labels such as horizontal, distributed, or collective leadership, shared leadership theory views leadership as a set of actions, rather than a designated role. Rather than an appointed leader "projecting downward influence on individuals, shared leadership entails the process of shared influence between and among individuals" (Pearce and Sims 2000, p. 116). Shared leadership reflects a web of mutual influences and shared responsibility and is associated with enhanced outcomes in a variety of settings including work groups, virtual teams, and virtual collaborations (Hoch and Kozlowski 2014, Perry et al. 1999, Sutanto et al. 2011, Wang et al. 2014). Likewise, we argue that leadership in online communities also emerges through interactions. Distinctively, though, the combination of open, voluntary participation and the paucity of formal leadership roles in online communities means that leadership is inherently shared. Whereas in formal organizations the relative concentration or distribution of leadership may be considered a strategic choice, we argue that shared leadership is an intrinsic property of online communities.

Finally, we consider the communication as constitutive of organizing theory as a pertinent perspective to understand online community leadership (Cooren et al. 2011, Taylor and Van Every 1999). This theory emphasizes the dynamic processes of communication in organizations and how these communication flows enact the social structure via interactions. Organizations are both a network of conversations and the symbolic dimension to interpret these conversations. These communicative interactions act as a structuring process for organizational processes and reinforce organizational processes (Robichaud and Cooren 2013). From this perspective, collaboration or even leadership cannot be conceived as independent of the text that forms the base of organizational conversations. These conversations build on the textual corpus to transcend the text to move to the realm of action and interactions. For example, when confronted with a text produced elsewhere in the organization, people evaluate it for relevance to their own context. They interpret it based on their own experience, and their reactions are shaped by norms within their specific community of practice (Taylor and Van Every 2010). This approach has been applied to online settings, to identify communication patterns of online leaders (Huffaker 2010, Zhu et al. 2012). Although the CCO framework appears most pertinent to the structured world of within-organization communication, its emphasis on explaining how conversation cycles support networking and social structuring makes it relevant to examine the utterances of online community leaders.

As our theoretical review of the four leadership theories indicates, significant differences exist between theories developed for explaining team and organizational leadership and the setting of online communities. We share the emphasis of functional leadership theory on the functions of leadership rather than the behavior of formally designed leaders. We draw on LMX theory to stress that leadership is contingent and situated. In agreement with theories of shared and distributed leadership, we recognize that leadership is not restricted to designated leaders. Finally, we draw on communication as constitutive of the organizing framework to emphasize the role of online interactions in sustaining leadership.

Applying Leadership Theory to Online Communities

Given that no single theory of leadership seems uniquely suited to online communities, we propose that multiple theories can be productively applied to this setting. Three major attributes of online communities necessitate adaptation of existing organizational leadership theories. First, like other voluntary collectives, there are few participants with formal power. Rather than formal roles and responsibilities dictating interaction and communication norms, efforts are predominantly performed in informal voluntary roles defined by behavior (Butler et al. 2007, Collier and Kraut 2012). When they exist, positions of formal power (such as moderation) appear to possess a limited range of rewards and sanctions. There are no tangible resources to distribute and few formal sanctions short of removing content or members. Thus, governance and leadership structures are emergent and highly situated to each community's setting (O'Mahony and Ferraro 2007). Second, compared with formal organizations, online communities are dominated by bottom-up emergent processes rather than top-down centralized interventions. They are fluid as they morph and change their boundaries, yet retain their shape and basic characteristics (Faraj et al. 2011). Finally, asynchronous written communication in online communities is intrinsically limited compared to face-to-face interactions. Participants lack the broad range of verbal nuances, nonverbal cues, and physical status characteristics that enrich other forms of communication. Yet, the online space is a social field where participants select distinct strategies of participation, produce and evaluate each other's content, and pursue distinction and marks of status (Levina and Arrigara 2015). Thus, online community members are constantly engaged in contribution strategies that positively differentiate them from others.

Given these unique characteristics of online community dynamics and membership, we suggest that any theorizing of online community leadership will require a contextualized synthesis of the four leadership theories described in the previous section. First, we must build on the functional leadership theory to evaluate the specific behaviors that differentiate leaders from nonleaders. Second, the online setting allows us to explore specific ties and interactions between leaders and nonleaders and thus offers a unique opportunity often unavailable in face-to-face settings. Third, given the size of the community and fluid membership, indications are that leadership is broadly distributed and thus would be shared. Finally, CCO theory, with its emphasis on how communication flows enact the social structure, is highly relevant for understanding online communities, where, by definition, one only "exists" if they post. The remainder of this section reviews existing empirical research on online community leadership to evaluate whether certain theoretical subtleties and empirical findings can further enhance our theorizing.

In a discussion of critical online community behaviors, Butler et al. (2007) identify four distinctive categories of maintaining infrastructure, social control and encouragement, external promotion, and content provision and consumption. Both infrastructure maintenance and social control require formal powers. Only authorized users can configure supporting communication infrastructure or remove unwanted content or members. These activities provide leadership through "a process of originating and maintaining the role structure" (Bass and Bass 2008, p. 18). In the moderated forums in our study, these activities are performed primarily by the designated roles of administrators and moderators. Other influential roles in building a community do not require formal powers. Any community member can provide social encouragement, promote the community externally, or create and read content. Nonetheless, some individuals will be more influential than others as they perform these activities. Applying the typology of leadership definitions described by Bass and Bass (2008), these participants can be said to emerge as leaders through influence processes resulting in recognition of informal leadership by others.

Recent research on leadership in online collectives has investigated the adjacent settings of open source software development, Wikipedia, and online communities. In a study of the governance structures and leadership in open source software development, O'Mahony and Ferraro (2007) collected interviews, secondary data, and project documentation to understand phases of governance over a 13-year period. Through an analysis of 815 participants during a time period of stabilizing governance they identified behaviors and characteristics that increased the likelihood of being assigned to a formal leadership role. They found that tenure, quality of contributions, and degree centrality all predicted leadership team membership. Likewise, Fleming and Waguespack (2007) performed a longitudinal analysis on 16 years of membership in the Internet Engineering Task Force to identify the types of human and social capital associated with moving into formal leadership roles in this voluntary community. They also concluded that network position (boundary spanning) and quality of technical contributions are associated with formal leadership roles. Although the communities in both studies have substantial in-person interactions, they nonetheless support the proposition that network position predicts leadership.

Multiple studies have also considered leadership characteristics in the production community of Wikipedia. Collier and Kraut (2012) analyzed 2,442 candidates under consideration for the formal leadership role of administrator in Wikipedia. The data, spanning six years of leadership deliberations, support the importance of network position to formal leadership advancement. In turn, Zhu et al. (2012) used a novel machine learning technique to evaluate the effectiveness of the leadership styles. Through the analysis of 1.6 million messages by 31,676 unique wiki editors, they found that leadership styles vary in effectiveness based on roles and that those in formal roles are more influential than other leaders. Together, these studies demonstrate the importance of formal roles in production-oriented communities as well as the value of considering structure and linguistics in relationship to leadership in online collectives.

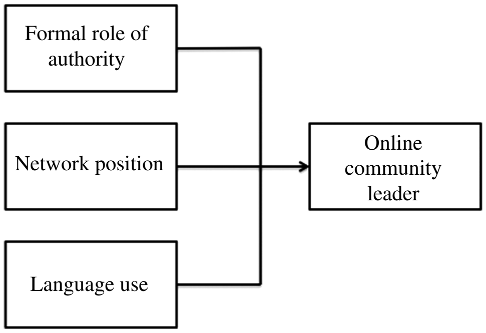
Huffaker (2010) took a linguistic analysis approach to develop a comprehensive look at leadership behaviors in online communities. Focusing on the leadership role of generating interaction, he analyzed the structural and linguistic characteristics of the participants whose posts have the most impact on community discussions. Based on two years of communication in 16 Google Groups (discussion forums), his study encompasses over 600,000 messages from over 33,000 participants. Controlling for communication frequency, he found strong support for the importance of multiple structural and linguistic measures in triggering replies, creating conversations, and diffusing group-specific language. Specifically, "online leaders influence others through high communication activity, credibility, network centrality, and the use of affective, assertive, and linguistic diversity in their online messages" (Huffaker 2010, p. 593).

In summary, our review of prior empirical findings lead us to conclude that leadership is influence based and involves aspects of all four theories described above. Thus, we propose an integrated model for leadership in online communities and adopt a popular definition of leadership (Yukl 2010): a leader is someone who is viewed by other participants as influencing what the online community does or how it does it. We further draw inspiration from a guiding idea of these studies: leadership is associated with participant roles and structural position. In the next section we describe a model of online community leadership that also incorporates the importance of language usage.

Model of Online Community Leadership

Our study builds on these previous empirical findings by identifying leaders through peer nominations and adopting a robust set of structural and linguistic measures. We investigate the characteristics of online community participants associated with exhibiting leadership. As shown in Figure 1, three characteristics of leadership stand out as most relevant to online communities and other open communication networks. First, participants in formally designated roles are more likely to be viewed as leaders. Second, filling an informal leadership role is not a single static designation but, rather, an emergent role based on the structure of repeated interactions. Third, not only is the behavior of regular interaction important but also communication qualities of those interactions matters.

Figure 1 Model of Leadership in Online Communities



Formal Roles of Authority

Many online communities, including those studied in this paper, grant formal authority to designated administrators and moderators. Typically, these participants have community-recognizable handles or identifying markers in their signatures, and as a result are likely to "have disproportionate influence, through possession of consensual prestige or the exercise of power, or both, over the attitudes, behaviors, and destiny of group members" (Hogg 2001, p. 188). There are three distinct processes that suggest those filling these formal roles will also be online community leaders. These relate to their recruitment, status characteristics, and role behaviors.

First, administrators and moderators are typically recruited from the most active and engaged participants. For an online community to grow from a handful of active users to hundreds, thousands, or more, the number of participants providing leadership must also grow (Butler et al. 2007). The most likely candidates for filling formal roles of authority are those who have demonstrated leadership qualities such as the interest and aptitude in helping to shape community dynamics. Given the importance of repeated interaction in establishing one's status online, active participation can be regarded as a sine qua non of being considered a leader.

Second, in the online communities in this study, the formally designated titles of administrator and moderator are prominently displayed next to all content posted by those users. With traditionally significant status characteristics (e.g., age, gender, ethnicity, personal appearance) largely hidden online, those that remain are even more salient (Hogg 2001). Finally, although these roles enjoy few formal capabilities, administrators and moderators play a major role in structuring interactions in online communities. Structuring of participant interactions is a leadership behavior (Reicher et al. 2005). Moderators and administrators can move and remove content, ban members, and use the threat of such to coerce desired behaviors. In summary, we propose the following:

Proposition (P1). *Occupying a formal role of administrator or moderator is positively associated with online community leadership.*

Communication Network Position

A central activity of online communities is written communication visible to all participants. Participant posting creates a communication network that is amenable to social network analysis. This approach has been applied both to the larger question of the structural role of leadership as well as to leadership in online communities. The emergent consensus is that leaders score highly in various centrality measures and also play a boundary spanning role to acquire information or resources (Balkundi and Kilduff 2006, Barge 1994).

In online settings, given the lack of face-to-face connection, communication network position is primarily based on where online contributions are made, how new ties are formed, and how those ties influence others' impressions (Dahlander and Frederiksen 2012, Donath 2007). Empirical studies indicate that online leaders tend to be longer-term participants of the group, entertain more ties with different others, and post frequently (O'Mahony and Ferraro 2007). Yet, in communities based on knowledge sharing, online leaders are not necessarily more "chatty" than others. In a study of a legal community, Wasko and Faraj (2005) found that experts, while being more central, were also suspicious of the validity of content provided by nonexperts and engaged in exchanges with little expectation of reciprocity. Taken together, these studies indicate the importance of a holistic view of leadership in communication networks.

For example, in online communities there is support for leadership being associated with network centrality, though results diverge on the type of centrality. A study of 16 Google Groups by Huffaker (2010) found that expansiveness (out-degree centrality) was associated with leadership behaviors, but brokering (betweenness centrality) was not. Looking at virtual collaboration supported by Second Life and in text-based chat rooms, Sutanto et al. (2011) found that both degree and betweenness centrality were associated with emergent leadership, but closeness centrality was not.

Closely related to centrality, the concept of core/periphery provides a complementary understanding of the structure of a communication network (Borgatti and Everett 2000). Compared with continuous measures of centrality, core/periphery suggests that there are distinct subgroups of participants with jointly occupied, structurally equivalent positions. Core/periphery structures have been identified in smoking cessation (Cobb et al. 2010) and video-blogging online communities (Warmbrodt et al. 2008). Membership in the core is associated with leadership in open source software developer communication networks (Crowston and Howison 2005) and in Wikipedia (Collier and Kraut 2012).

The online communities studied in this paper are supported by asynchronous discussion boards and are organized with participation structures of threads and forums. Some participants may focus their participation within a single topic, whereas others may have participation spanning the topic boundaries of threads and forums. The former have low boundary spanning, and the latter have high boundary spanning. Although high boundary spanning has been associated with leadership characteristics in multiple domains, including knowledge-intensive work (Levina and Vaast 2005) and open innovation communities (Fleming and Waguespack 2007), overall evidence of an association is mixed (Reagans and Zuckerman 2001). The primary value of boundary spanning is derived through information brokering (Burt 1995). In online communities where posts are visible to all members, the ability to broker information is reduced. Furthermore, our sample is of complex knowledge-rich topics where no single individual can be an expert in all areas. This further reduces the ability to broker information and favors participants who have deep, rather than broad, knowledge to share. Thus, we argue that participants who are central in the communication network, are part of the communication network core, or have low boundary spanning are more likely to be online community leaders than others. In summary, we propose the following:

Proposition (P2). *Communication network position (high centrality, a core position, and low boundary spanning) is positively associated with online community leadership.*

Language and Leadership

Both where and how communication occurs are salient to organizational processes. Indeed, many scholars have recognized that "communication is the medium through which leadership occurs" (Barge 1994, p. 29). For example, discursive leadership theory focuses on communication to understand behaviors consistent with leadership (Fairhurst 2007). Barge (1994) stresses that linguistics is integral to phrasing persuasive messages, yet should also be tailored to individual contexts. Barrett (2008) notes the importance of language as a way for leaders to influence others and recommends use of concise positive messages. In regard to leadership in online communities, these works suggest that it is not just a matter of which participants communicate with each other, but also the characteristics of that communication.

Looking more closely at communication online, multiple studies find an association between language usage and demonstrating leadership. Yoo and Alavi (2004) analyzed communication among team members of seven executive student project teams. They found that the team members that emerged as leaders wrote longer and more frequent emails than other team members. Wickham and Walther (2007) analyzed discussions of 18 small groups working on a decision-making task. They also found that higher levels of communication activity were consistent with being identified as a group member exhibiting leadership. In a review of leadership perceptions in both online and offline small group settings, Hollingshead (2011) notes that the quantity of participation is highly correlated with leadership. However, her review also notes that in knowledge-oriented forums, the quality of participation is more closely associated with leadership than merely quantity of participation. Frequency alone is not enough to be a leader; the quality of communication also matters. Additional studies analyze behaviors in terms of leadership styles and language characteristics. For example, Huffaker (2010) found that participants identified as online community leaders used more affective and assertive language than others. Finally, in a study of influence behaviors seen in Wikipedia, Zhu et al. (2012) identified different types of language used as associated with different leadership styles. Together, these studies support the general idea that leadership is associated with differences in language use in online settings, but provide limited guidance regarding specifically how those differences manifest themselves.

We propose a linguistic analysis model to shed further light on language use consistent with being considered a leader in online communities. Drawing on work in computer science, artificial intelligence, and linguistics, natural language processing (NLP) offers a promising approach to enable computers to derive meaning from human utterances, and more crucially, to derive a multidimensional understanding of bodies of text (Clark et al. 2010, Mitkov 2005). The NLP approach splits language analysis into theoretical (often hierarchal) levels (Mitkov 2005). Our operationalization of the NLP algorithm relies on generating data along these four major dimensions of language: morphology, lexicography, syntax, and semantics.

Several organizational leadership theories support the relevance of language to leadership. Empirical evidence from written electronic communication finds that leaders use language differently than nonleaders. NLP provides a systematic approach to identifying those differences. We propose five core linguistic features, each mapping to one of the four major dimensions of language, as consistent with online community leadership: readability (morphology), vocabulary richness and external linking (lexicography), prototypicality of vocabulary (syntax), and positive sentiment (semantics).

First, we propose that text readability is positively associated with leadership communication. Brevity and succinctness are characteristic of effective communication (see Zinsser 2006), and conciseness is recommended to achieve a leadership purpose (Barrett 2008). Holding all else equal, an online participant who can express their ideas simply (with improved readability) is more likely to influence others than one who expresses their ideas in a difficult to read manner.

Readability and vocabulary are related, yet distinct, linguistic features. Whereas readability is based on characteristics of single words (e.g., length, number of syllables) and sentences (e.g., number of words), vocabulary richness reflects how many different words someone uses. A body of text containing a large number and variety of short simple words is more readable with more vocabulary richness than text with a small number of long, complex words. All else equal, someone who commands a larger vocabulary has more tools available to word and reword ideas, thus to better influence others. Huffaker (2010) in his study of 16 Google Groups measured online leadership as having influence on the communication behaviors of other group members. Participants with increased linguistic diversity had more influence than those with less linguistic diversity.

Providing direct access to online resources (via hyperlinks) is another aspect of online written communication. Providing URLs in online text is likely to increase online influence for multiple distinct reasons. First, providing a link can serve as verifiable evidence for arguments made in online rhetoric. Second, a link may provide a resource of value to the online community, be it news, information, or entertainment. Third, a link may directly address a question or concern of others. These are three pathways of influence that may demonstrate leadership.

Another linguistic feature related to vocabulary is how distinctive word choices are in relationship to the frequency of words used by others. In an online community this takes the form of comparing all of the words used by a single individual to all of the words used by the rest of the community. The closer an individual's word usage is to the collectives', the more they represent an average or prototypical vocabulary use for that collective. In social identity theory, prototypicality is both a process leading to as well as an outcome of social influence and leadership (Ashforth and Mael 1989, Hogg 2001). Thus, we expect that the more prototypically a participant uses the vocabulary of an online community, the more likely they are to be identified as an online community leader.

Finally, we consider the role of sentiment in communication. Sentiment is commonly associated with leadership. Leader mood is contagious (Sy et al. 2005). Positive emotions create affective bridges that serve as channels of influence, and some scholars view that "shared affect could be more salient basis for group formation than shared cognition" (Weick 1969, p. 14). Leadership emerges from positive sentiment and microeffective events (Johnson and Dasborough 2008). In summary, we propose the following:

Proposition (P3). *Unique patterns of language use (readability, vocabulary richness, external linking, prototypicality of vocabulary, and positive sentiment) are positively associated with online community leadership.*

Research Method

Research Design

Testing the propositions requires four different types of data. First, survey data are used to measure the dependent variable of online community leadership. Participants in three communities focused on technical topics were asked to identify other participants who they regarded as most influential in what that online community did or how that online community did it. Second, the participants in formal roles of authority (administrator or moderator) were collected from public lists posted at each of the three online communities. Third, the structure of the communication network was gathered through automated collection that identified how participants interacted through threaded discussions. Finally, the full text of all of the parsed posts was collected to document the content of interactions. This text forms a corpus of full text messages analyzed with NLP algorithms. Because the focus of this study is to compare participants (online community leaders versus other participants), measurements are aggregated to the participant level.

Table 1 provides an overview of the targeted communities, all of which focus on technical topics and use vBulletin, an open, asynchronous, web-based message board technology. Community discussions are organized by message thread, with each message thread belonging in a single higher-level topic forum. These communities were chosen from a sample of several dozen online communities surveyed in the spring of 2008 for a broader-based study of online participation (citation blinded). These three were randomly selected from those with at least a dozen survey-nominated leaders and a year of presurvey, full message-level communication available.

Identifying Online Community Leaders

To test the propositions, we needed to identify participants who are online community leaders. Most existing empirical studies of leadership behaviors outside of formal positions of authority focus on small work teams in educational or organizational settings. As such, they predominately operationalize leadership through work team peer ratings, where each participant rates the leadership qualities of all other team members (Pfeffer and Cialdini 1998, Walter et al. 2012). This approach is not practicable in online communities with thousands of active participants. Instead, online community leaders were identified by asking survey respondents to name up to three participants who had the most influence on what the communities does or how it does it (wording directly applied from Yukl's (2010) definition of leadership). The terms "leader" and "leadership" were intentionally avoided in the prompt so that respondents would focus on leadership outcomes rather than formally designated roles. Consistent with our theoretical stance that online community leadership is a local temporary designation, we consider anyone identified

Table 1 Online Community Descriptive Statistics

Online community	Blender Artists	Gearbox Software	Northern Sounds
Tagline	Community of artists using Blender, a 3D creation tool	The official community of Gearbox games	Northern Sounds software
Collection URL	blenderartists.org/forum	gbxforums.gearboxsoftware.com	northernsounds.com/forum
Inception	October 14, 2001	July 12, 2002	February 1, 2003
Members	10,264	1,644	2,488
Forums	28	27	37
Threads	32,656	5,383	8,472
Posts	308,682	118,924	51,472
Words	17,088,714	4,673,512	4,357,698
Survey responses	19	16	21
Participants in formal roles	8	11	8
Online community leaders	23	21	15

Table 2 Relationship Between Identified Online Community Leaders and Participants in Formal Roles

	Blender Artists	Gearbox Software	Northern Sounds	Total
(A) Identified online community leaders	23	21	15	59
(B) Participants in formal roles of authority	8	11	8	27
(C) Identified online community leaders also in a formal role of authority	4	5	3	12
% of participants in formal roles (B) who are also online community leaders (C)	50% (4 of 8)	45% (5 of 11)	38% (3 of 8)	44% (12 of 27)
% of online community leaders (A) who are also in a formal role (C)	17% (4 of 23)	25% (5 of 21)	20% (3 of 15)	20% (12 of 59)

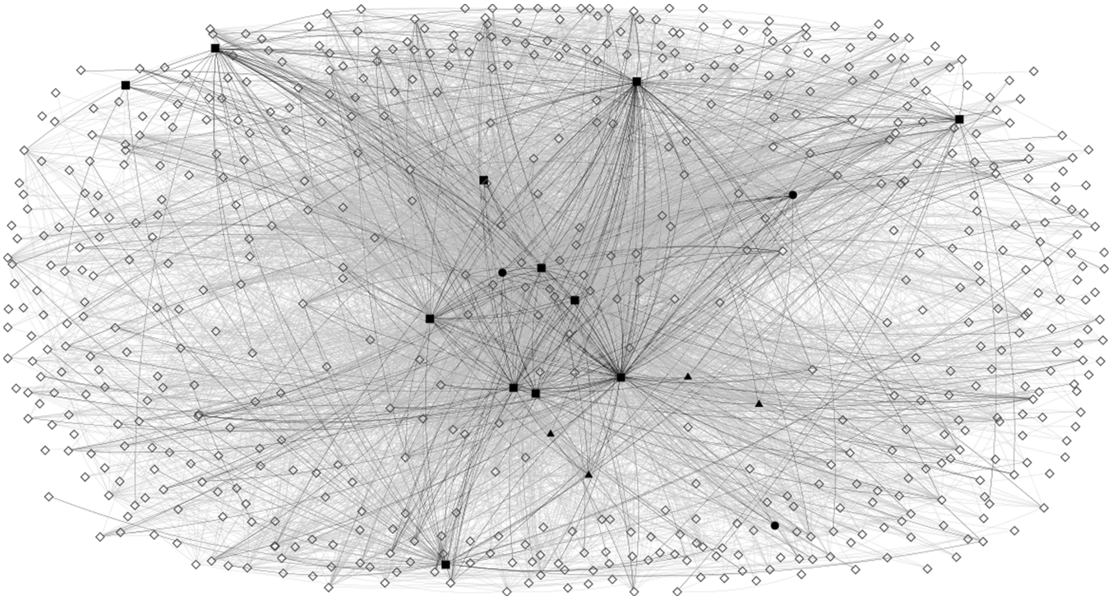
by a fellow participant as demonstrating leadership to indeed be an online community leader. As such, the dependent variable in our analysis is an ordinal value reflecting whether any other participants have nominated the focal participant as a leader.

Identifying participants occupying formal roles of authority is a straightforward process. The three online communities in our sample each display a public list of moderators and administrators. This list was captured immediately prior to the survey response period. As demonstrated in Table 2, participants identified as online community leaders and participants with formal authority are related yet distinct categories. Consistent with the first proposition (above), 44% of those in the latter role are also in the former (50%, 45%, and 38%, respectively, in the three communities). The validity of the online community leadership construct is strengthened by the much smaller overlap between the two categories. Only an average of 20% of identified online community leaders are moderators or administrators (17%, 25%, and 20% in the three communities). The two measures have a 0.33 correlation ($p < 0.001$) in the analyzed sample. (Correlations are provided in the appendix for all of the study measures.)

Structural Model

Archival data were used to calculate the communication network structure measures of centrality, core position, and boundary spanning. Communication in the studied communities was modeled as affiliation networks with two node types: participants and threads (Borgatti and Halgin 2011). A communication network link (i.e., an edge in the graph) exists between a participant node and a topic node when the participant posts to a message thread. The network link carries the attribute of the communication such as the text, date of the post, and order of the post in the discussion. Figure 2 shows a graph representing the relationships among active participants (top 20% of frequent participants). Black circles represent online community leaders who are also in a formal role of administrator or moderator. The black squares are other online community leaders. The black triangles are administrators and moderators not identified as leaders. White diamonds represent all other participants.

Figure 2 Communication Among Active Participants of Northern Sounds



Dark edges represent communication originating from community leaders, whereas light edges represent all other communication in the community.

Several items of note can be seen in this representation of the active core of this community. First, a central network position is neither necessary nor sufficient for being an online community leader. Leadership is a local designation; within the core leaders occupy both central and peripheral positions. Second, there is no discernable relationship between serving in a formal role of authority and network position. Although a relatively high percentage of those in formal roles are also identified as online community leaders, network position does not suggest which ones. Third, although a small fraction of the community, leaders communicate with many other participants regardless of their formal designation or network position. This is visually evident by the spread of darker edges over the community.

Modeling the community as a communication network also allows for the investigation of network characteristics associated with nodes and allows us to numerically ground the observations drawn from examining the graphical representation (Knoke and Yang 2008). Several measures exist to describe centrality in a network. The first is degree centrality, the fraction of nodes in the graph connected to the node under consideration. The second is betweenness centrality, which is the sum of the ratio of all pair's shortest paths that pass through a focal node. The third is closeness centrality (measured as the reciprocal of the normalized average distance to other nodes), a measure of how long it takes to sequentially disseminate a message to all other nodes in the network. Because these three network measures were highly correlated in our sample, only one was retained for the measurement model. Betweenness centrality was chosen as the most theoretically meaningful of the three measures because it takes into account both the local connections of a node as well as its global position in the network (Mehra et al. 2001). Supporting this approach, post hoc analysis using alternative measures of centrality did not affect our results.

We also use core/periphery measures to account for the possibility that the network had an active core of highly involved participants primarily engaged with each other and a larger periphery of less involved (peripheral) participants. Indeed, studies of online communities show that the behavior of participants is impacted by their position vis-à-vis the core (Dahlander and Frederiksen 2012, Liu 2011). The k-core number is used to divide the network into layers of cores. The cores are subnetworks with k connectivity. For example, it is possible to fragment the 1-core by deleting one edge, whereas at least two edges need to be deleted to fragment the 2-core. A higher k-core number occurs for nodes that are located in densely connected parts of a network (Seidman 1983).

Boundary spanning is operationalized as a ratio of the number of messages to the breadth of areas those messages appear in. Specifically, boundary spanning is measured as the ratio of the number of unique threads a participant posted messages in, divided by their total number of posts. Lower values of this measure occur when a participant concentrates their posts into a smaller number of threads. Higher values occur when a participant posts messages in many different threads. Thus, this measure reflects the degree of specialization of topics. Together, centrality, core position, and boundary spanning provide a robust structural assessment of position in the communication network.

Model of Language Usage

Advances in both online data availability and computing processing speed have opened up new opportunities for automated communication analysis. The application of both NLP and computation linguistics (CL) has recently flourished (Clark et al. 2010, Jurafsky and Martin 2008), and both are well suited to help assess how subsets of a group, such as leaders and others, differ in language usage. NLP and CL are used in real-life applications, including speech recognition, text translation, and question answering, that exist in a wide variety of platforms ranging from mobile devices such as the Siri personal assistant on the iPhone (Aron 2011) to supercomputers such as IBM Watson, the world champion of Jeopardy (Ferrucci 2010). The ultimate goal of NLP is to mathematically model the understanding and generation of human language.

In this paper, we apply NLP algorithms to better understand how language usage is associated with online community leadership. Table 3 provides a representative sample of posts from the Northern Sounds community by different participant types. Examining these posts gives a rudimentary idea of how leaders may differ from nonleaders in terms of language use. Online community leaders tend to use simple positive language with high readability. Other active participants tend to use less familiar language with more complex sentences. We expand on previous research identifying the importance of language usage (Hollingshead 2011, Huffaker 2010, Zhu et al. 2012) by applying a systematic computational approach to quantify online community participant messages.

Given the complexity of human language in action, the NLP approach splits language analysis into theoretical (often hierarchal) levels (Mitkov 2005). Our operationalization of the NLP algorithm relies on generating data along these four major dimensions of language: semantics, syntax, lexicography, and

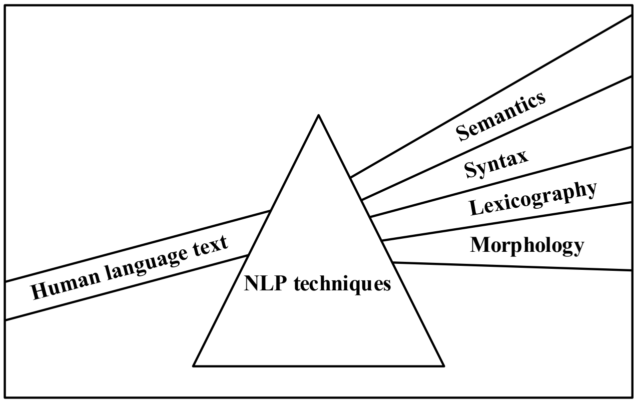
Table 3 Example Posts in Northern Sounds Online Community by Participant Type

Participant type	Representative post text	Characteristics of text
Online community leader in formal role of authority	"'Nieves, ….' Excellent rhythms going on in this piece. Excellent percussion writing and full of energy. I agree with Reegs that Drumlines would like this. I can't wait to see what you can do with the Marching Band library. Keep on doing what you are doing."	Simple positive language with high readability
Online community leader not in formal role of authority	"Thank you very much you all good people, who made this course possible. It was a great experience."	Simple positive language with high readability
Active participant in formal role of authority	"I have been a member of this community for a while now but this is the first time I have posted a work in the Listening Room. http://www.michaelsroom.co.uk/Handel—Organ Concerto in F (The Cuckoo and the Nightingale) 2nd Movement.mp3 For those who are unfamiliar with this, the nickname (Cuckoo and Nightingale) comes from the bird song motifs to be heard in this movement. Coincidentally, this work was completed by Handel in April (2nd) 1739."	Less typical language for the community
Active participant not in formal role of authority	"In my opinion, wait a few months and see how GS4 turns out and let the bugs get fixed, then its time to get 1 machine that is as bad to the bone as you can afford (quad core, 8 gigs ram?) with a 64 bit OS (xp64 or vista64) and upgrade to GS4. I think once the kinks are worked out, and knowing tascam there will be some lol, that it will be quite awesome to go with a beefy GS4 machine."	Lower readability

morphology. Together, these measurements allow the analysis of how language is used in online communities using multiple and complementary linguistic perspectives (a prism model). Like a prism breaking light into its full spectrum, NLP analytical techniques (Bird et al. 2009) break text into multiple components (Figure 3). Identifying a full spectrum of linguistic characteristics provides a robust method to compare leaders and other online community members based on their expressed language corpus. The starting point for our analysis is morphology (the subword level) and continues to semantics (the meaning of text). The goal is not to completely model the use of language at each intervening level, but rather to identify representative indicators to compare participants of online communities. Each of the four levels (morphology, lexicography, syntax, and semantics) is described further below.

Morphological Analysis. Morphology studies how words are formed in natural language. More precisely, it is how the words are segmented into components that form those words via concatenation (Goldsmith 2001). Two main types of such decomposition exist: morphophonology, in which the subcomponents correspond to spoken syllables, and morphosyntax, in which the subcomponents are syntactical (such as prefixes and suffixes). An example measure is the number of syllables per word. Words with more syllables are considered more complex; their usage indicates a higher command of language (Gunning 1969). At the morphological level, there are three well-known indices of readability: the Automated Readability Index (ARI), the Flesch–Kincaid Reading Ease test (Kincaid et al. 1975), and the Gunning–Fog Readability Index (Gunning 1969). Because all three measures are highly correlated in our data set, we choose the simplest of the three, the ARI, for analysis. The ARI takes into account the number of characters, words, and sentences in a post. It yields a higher score for longer words and longer sentences. It is computed with the following equation:

Figure 3 The Natural Language Processing Analysis Prism



$$\text{ARI} = 4.71 \cdot \frac{\#characters}{\#words} + 0.5 \cdot \frac{\#words}{\#sentences} - 21.43.$$

Lexical Analysis. Whereas morphological analysis focuses at the subword level, lexical analysis focuses on characteristics of participants' vocabulary at the word level. For example, the number of words used by an online community participant can be assessed as well as the qualities of those words (e.g., by matching them against precompiled dictionaries). First, a dictionary is compiled for each participant. The dictionary contains the unique words that the participant used in their posts. Next, the size of a participants' dictionary is normalized based on their number of posts. The resulting indicator (vocabulary richness) measures the richness of vocabulary. In addition, the use of hyperlinks is considered a special vocabulary. The average number of links per post is also calculated for each participant.

Syntactic Analysis. The syntactical level examines how words are combined and used to form sentences and posts. This paper adopts an NLP technique called statistical language modeling (Jurafsky and Martin 2000) that assigns a probability to a sentence in a textual corpus given the likelihood of that sentence based on the rest of the textual content of the corpus. This probability is an indicator of whether that sentence conforms in its syntax with the rest of the sentences in the corpus. For example, if we take the Bible as a corpus, a sentence like "Computers are used to process words" will have a very low probability, whereas a sentence like "The word was with God" will have a high probability. The latter sentence is thus more prototypical of the Bible than the former.

Prototypicality is measured as the inverse of entropy, a sophisticated approach to comparing individual and group language usage. Statistical language models (Jurafsky and Martin 2000) allow for calculating the likelihood that a word, sentence, or set of sentences is representative of the language used in a bigger collection of text, also called the text corpus. As such, it can be used to compute the probability that a post was authored by a participant of the community given what all other participants wrote.

Building a model that estimates the exact sentence probability is computationally challenging because of the large number of potential word combinations in sentences of varying length. The solution is to approximate the computation of the sentence's probability by considering word sequences of fixed length in the sentence. Those sequences are called N-grams. A sequence of one word is called a unigram, a sequence of two words is called a bigram, and a sequence of three words is called a trigram. In most cases, a trigram model ($n = 3$) provides a strong approximation and has been considered the standard statistical language model for more than 30 years (Clark et al. 2010). A trigram statistical language model is used. For ease of exposition, a bigram model representation is shown in Table 4.

The probability of a two-word sequence (i.e. a bigram) is estimated by how many times the two words occur together in the text corpus divided by the frequency of the first word (Step A in Table 4). The probability of a complete sentence is calculated as the multiplication of the probability of the sequence of bigrams in the sentence (Step B). Because the probability of a sentence is a very small number and because this number depends on the length of the sentence, the entropy of a sentence is defined as the negative log of its probability divided by the number of words in the sentence (Step C).

Table 4 **Computing the Probability of a Sentence W Composed of N Words Using a Bigram Model**

(A) Probability of a bigram	$P(w_n \mid w_{n-1}) = \text{Count}(w_{n-1}w_n)/\text{Count}(w_{n-1})$
(B) Probability of a sentence	$P(W) = P(w_1 w_2 \ldots w_n)$ $= P(w_2 \mid w_1) \cdot P(w_3 \mid w_2) \cdot \ldots \cdot P(w_n \mid w_{n-1})$
(C) Word entropy of a sentence	$\text{Entropy}(W) = -\log(P(W))/N$

The entropy number can be used to compare two sentences in light of the training corpus. The sentence with higher entropy has a lower probability of occurrence and is (probabilistically) more unique than the first one, whereas that of lower entropy has a higher probability of occurrence and is (probabilistically) more of an average sentence than the first one. Bringing this measurement to the level of the participant, participants whose posts have higher entropy on average are contributing unique posts that deviate from what other participants are writing, and vice versa. Thus, participants whose contributions are characterized by high levels of entropy are not prototypical of the conversation of the community and are possibly offering novel information. As such, prototypicality is measured as the additive inverse of entropy.

Semantic Analysis. At the level of semantics, the goal is to go beyond the structure of words and sentences to identify the meaning of what is written. As such, the semantic analysis is the first step of natural language understanding—a step that is considered the most challenging in linguistic analysis (Shahaf and Amir 2007). A simple form of semantic analysis is assessing the sentiment of posts in terms of their polarity (i.e., positive versus negative) (Pang and Lee 2008). For our study, the sentiments expressed in posts were assessed in terms of their negative versus positive polarities. A word-based classifier of sentiments based on a dictionary of emotionally rated English words, AFINN (Nielsen 2011), was used. The dictionary used is a customized version of the package ANEW (Bradley and Lang 1999), which provides a set of normative emotional scores for a large set of English words. AFINN customized the ANEW dictionary to tailor it to the Internet language of Web logs, discussion forums, and tweets (Nielsen 2011). In addition, AFINN associates a score to each post based on the emotions expressed within the words of that post. A negative score implies negative emotions or polarity in the post, and vice versa for a positive score. A score of zero implies a neutral tone.

A number of additional preprocessing and data-cleaning steps were required to facilitate linguistic analysis of participants' postings. First, all of the collected posts were preprocessed to remove HTML formatting (e.g., bold, italics). Second, the special content of Web links and formal quotes of other participants were identified and removed. These filtering steps are important (a) to focus the language modeling toward what a participant says rather than the text that is being repeated from a previous post and (b) because many of the linguistic measures are poorly suited to marked-up text. Finally, the linguistic characteristic of each post was assessed with measurements aggregated per online community participant. As described above, the four dimensions of language process focused on are morphology, lexicography, syntax, and semantics.

Model Testing

Because online community leadership is a binary categorical variable, logistic regression (Long and Freese 2006) is a natural choice to model its relationship with the independent variables (summarized in Table 5). Since participants are nested within communities, we use a random-effects logistic regression model where the effect of community membership on an emergent leadership role is a random-effects coefficient. Three potential concerns arise regarding this analysis approach: First, the dependent variable is sparse (i.e., very few of the participants were nominated as leaders). Second, most of the structural independent variables are not normally distributed. For example, the centrality and core variables follow a power law that is commonly found in network data (Faraj et al. 2008). Third, because of the power-law distribution and the large sample size, many outliers are found in most variables. The three concerns are interrelated and are indeed typical characteristics of large networks.

The sparseness of data could affect the power of the statistical analysis, but this is addressed by the large sample size (thus providing greater statistical power). However, the large sample size poses its own concerns. The large data set we obtained (14,396 participant observations across the three communities) can be criticized from both a theoretical and practical perspective. First, from a theoretical perspective, communication networks exhibit a power-law distribution structure (Faraj et al. 2008). Most participants contribute little while a few contribute a lot to the community. Indeed, 40% of participants in the three communities contributed only one or two posts. Therefore, it is problematic to compare leaders to everyone else knowing that most participants contribute few posts. A more appropriate comparison group is community participants who are also frequent contributors but were not nominated as leaders. Second, from a practical perspective, the large sample size leads to statistically significant results. The ability to harvest large data sets from the Internet may be problematic when practical interpretation of significant coefficients is difficult (Royall 1986).

To address all three issues in a rigorous way, we focus on the most theoretically and computationally relevant subset of the available sample. Because we are interested in comparing leaders to participants who are equally engaged and contribute to the community but were not identified as leaders, we use the number of messages a participant posts to the community as a threshold variable. This variable follows a power-law distribution. We keep observations of participants with their number of messages in the top 20th percentile. This corresponds to more than 18, 64, and 14 messages in Blender Artists, Gearbox Software, and Northern Sounds, respectively. The new sample size is reduced to 2,947 observations from the original 14,396 observations. Finally, we use robust standard error estimation to deal with the misspecification of the normal distribution in the independent variables. (Variables with a power-law distribution remain so in the reduced sample size because the power-law distribution is scale free.) Descriptive statistics and correlation tables (Tables A.1–A.14) are provided in the appendix.

In summary, we use a hierarchical analysis technique to test the extent of association between our hypothesized variables and online community leadership. The first model analyzes the association between leadership and formal roles of authority. The second

Table 5 **Participant Measures**

Measurement	Type	Description
Online community leader	Leadership	Identified as a leader by a fellow participant
Formal role of authority	Leadership	In a formal role of authority (administrator, moderator)
Centrality	Structural	Betweenness centrality of participant (larger value is more central)
Coreness	Structural	*k*-core number of participant node (larger value is more in the core)
Boundary spanning	Structural	Ratio of number of unique message threads posted is divided by total number of posts.
Readability	Linguistic (morphology)	Automated Readability Index
Vocabulary richness	Linguistic (lexicography)	Vocabulary richness (average number of unique words per post)
External linking	Linguistic (lexicography)	Average number of Web links
Prototypicality	Linguistic (syntax)	Prototypicality of participant language use when compared to other participants of the same community
Positive sentiment	Linguistic (semantic)	Average sentiment polarity score

model analyzes the impact of network measures in combination with formal roles. The final model adds linguistic variables. All three models are random-effects logistic regressions. Finally, to more directly interpret results, we standardized all of the research variables except the two (binary) leadership variables.

Two types of evaluations are employed to compare the three models. First, we look at the independent variables' coefficients and goodness-of-fit indices in the three models to evaluate the effect of these variables on leadership in the studied communities. Second, we perform two model difference tests comparing the nested models to judge the added value of linguistic variables in determining leadership. We also perform postestimation tests for the full model (C) to ensure the validity of the results. We test for multicollinearity using variance inflation factors, and we test for overfitting using cross validation. Results of these additional validation steps are reported in the appendix.

Results

The analysis results are provided in Table 6. Examining the regression coefficients of participant-level variables, a formal role is the most important predictor for online community leadership. Holding everything else constant, a participant who is in a formal role of administrator or moderator is 35 times more likely to be viewed as a leader than other active participants. Structural and linguistic variables are also important. For example, increasing centrality by one standard deviation increases the odds of leadership by 150%. Similar increments in coreness, readability, prototypicality, and positive sentiment all enhance the chance of being identified as a leader. However, an opposite effect is found for boundary spanning and vocabulary richness: a one standard deviation increment in boundary spanning or vocabulary richness almost halves the odds of being identified as a leader. Only external linking turns out to be nonsignificant in predicting online community leadership.

The intragroup/intraclass correlation reflects the effect of group membership on being identified as a leader. Because identified leaders are equally distributed among groups (Table 2), the intragroup correlation is on the low side (9% in Model C). This is an important indicator because it suggests that group membership does not overshadow participant-level variables in predicting leadership. To evaluate the added value of structural variables and linguistic variables, we perform two χ^2 difference tests comparing Models A and B and Models B and C. The two tests are significant, indicating that both structural ($\Delta\chi^2 = 99.89$, $\Delta df = 3$) and linguistic characteristics of participants ($\Delta\chi^2 = 19.34$, $\Delta df = 5$) are important predictors of online community leadership.

Table 6 **Hierarchical Logistic Regression Model with Dependent Variable of Online Community Leadership**

Measure	Model A	Model B	Model C
Group-level coefficients			
Intragroup correlation	0.11	0.37	0.096
Participant-level coefficients as odds ratios; standard errors in parentheses			
Formal role of authority (P1)	58.98^{***}	37.45^{***}	35.20^{***}
	(28.00)	(20.16)	(19.86)
Structural measures (P2)			
Centrality		1.58^{***}	1.54^{***}
		(0.12)	(0.14)
Coreness		3.90^{***}	2.59^{**}
		(1.44)	(0.87)
Boundary spanning		0.50^{***}	0.57^{**}
		(0.10)	(0.12)
Linguistic measures (P3)			
Readability			1.33^{*}
			(0.19)
Vocabulary richness			0.45^{*}
			(0.15)
External linking			1.21
			(0.16)
Prototypicality			1.66^{*}
			(0.35)
Positive sentiment			1.51^{*}
			(0.28)
Goodness-of-fit indices			
Log likelihood	−248.5	−198.6	−188.9
χ^2	73.75	115.4	118.6
Akaike information criterion (AIC)	503.0	409.2	399.8
Bayesian information criterion (BIC)	521.0	445.1	465.7
Comparison with previous model ($\Delta\chi^2$)		99.89^{***}	19.34^{**}

Note. $N = 2{,}947$.
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

In summary, both structural and linguistic participant-level variables are associated with leadership in addition to formal roles of authority. On the structural side, a more central position toward the core of the community increases the odds of a participant being identified as a leader. The opposite is true for boundary spanning; a participant who spreads their messages across different threads and topics is less likely to be identified as a leader. On the linguistic side, a participant with language that is more readable, has a simpler vocabulary, is more prototypical, and is of more positive sentiment is more likely to be identified as an online community leader.

Model Validation

To ensure the validity of the analysis, we employed several postestimation tests. Multicollinearity is of concern because of the potential overlap among related measures used to evaluate structural and linguistic characteristics. Although standardizing the independent variables can alleviate multicollinearity in data (Barry 2011), we also tested for this post hoc. We computed the variance inflated factors (VIFs) of the regression variables after estimation (results in the appendix). All factors are below 2 with an average of 1.31, indicating the absence of multicollinearity concerns.

Next we address the concern that our model overfits the research data and could have little validity outside the research setting. This is a typical issue in machine learning and classification tasks because of the existence of noise in the training set making it difficult to judge whether the model parameters fitted the real data or the noise (Hart et al. 2001). We partially addressed this issue by reducing our data set to 20% of the original sample size and focusing only on participants with a large number of messages comparable to those of leaders.

As further validation of the model's robustness, we conducted an area under the receiver operating characteristic (ROC) curve analysis and computed a 10-fold cross-validation analysis (Hosmer and Lemeshow 2005, Kohavi 1995). The area under the ROC curve analysis is most suitable for testing the ability of two-class classifiers to detect the true signal and separate it from noise (Hosmer and Lemeshow 2005). The area under the curve ranges from zero to one, with any value above 90% indicating an excellent discriminative ability. Our model achieved an area of 91.81%.

Next, we checked external validity by conducting a 10-fold cross-validation test (Kohavi 1995). The measurement sample was split into 10 randomly selected subsamples, and the following procedure was repeated for each of those 10 samples: (1) the measurement model was run for a subsample; (2) the value of the dependent variable in the remaining 9/10 of the data was predicted from the regression coefficients calculated in the subsample analysis; (3) the root mean square error (RMSE) was calculated as a measure of the difference between the predicted values and the actual values (for the 9/10 subsample). As noted in the appendix, the average RMSE of these 10 analyses is 12%, indicating good external validity. Because of the two-class setup, this RMSE value corresponds to 90% accuracy of predicting the classes of instances outside of its training set correctly (Alpaydin 2004). Although other learning algorithms (such as a naïve classifier) may achieve a better RMSE, taking into account that other goodness-of-fit indices were also good (Table 6), the model indicates good external validity (Wolpert and Macready 1997).

Sensitivity Analysis

We performed several additional tests both to assess the sensitivity of results to the analysis sample selection and also to further explore how different subsets of online community leaders compare to one another. Additional logistic regressions are provided in the appendix for (a) the full data set and (b) the analysis data set with all participants in formal roles of authority removed. Support is found for all three propositions in both of these tests. This strengthens the conclusion that a formal role of authority, structural characteristics, and linguistic characteristics are all associated with being identified as an online community leader.

The sensitivity analysis also provides more nuance regarding individual structural and linguistic characteristics. Analysis of variance tests are reported for (a) a comparison of online community leaders with and without formal roles of authority and (b) a comparison of online community leaders identified by one participant to those identified by two or more participants. No significant differences were found in either structural or linguistic variables between online community leaders with and without formal roles of authority. Differences are noted, though, between leaders identified by a single participant and those identified by two or more participants. In terms of structural variables, online community leaders identified by two or more other participants post more and have higher centrality compared to those identified a single time. Online community leaders identified by two or more other participants have lower vocabulary richness and higher prototypicality compared to those identified a single time. Using simpler language that is most familiar to the participants is consistent with the highest likelihood of leadership identification.

Discussion

The goal of our research was to investigate whether, beyond network position, online community leaders had distinctive written communication patterns. Using participant surveys to identify leading online community members, this study analyzes a year of communication network history and message content to identify whether leader contributions have unique qualities compared to the utterances of other core community participants. We first examine how communication network position–in terms of formal role, centrality, membership in the core, and boundary spanning–affects the likelihood of being seen as a leader. Then we contribute a novel use of textual analysis to develop a language model of utterances in the community to evaluate how convergent or divergent leader language is compared to that of the community as a whole. Our findings suggest that the most influential participants of any online community, those viewed as leaders by other participants, are not just among the most central, but also post a large number of positive, concise posts with simple language familiar to other participants. Thus, leadership is not merely filling an assigned role nor occupying a communication network position. Online community leadership is multifaceted, enacted through unique language patterns, and based on the perception of others.

Theoretical Implications

Our paper makes four major contributions to the understanding of leadership in online communities. First, our findings lend support to a multifaceted approach for understanding leadership in online communities. We integrate four sets of empirics (formal leadership roles, peer nominations, network position, and content of utterances) to offer a deeper understanding of leadership processes in online settings. Our findings build on and augment previous studies that had prioritized network position as a proxy for leadership (Huffaker 2010, Sutanto et al. 2011) by delineating the relative importance of network position compared to other correlates of leadership. In addition, by comparing leaders to other participants of equivalent rank, we were able to establish the ways by which leaders demarcate themselves from other members at the core of the community. Thus, we offer a more fine-grained evaluation of the activities of those active in the core and thus extend earlier findings regarding the core/periphery perspective on online communities (Cobb et al. 2010, Collier and Kraut 2012, Crowston and Howison 2005, Warmbrodt et al. 2008).

Second, our findings align with recent theorizing in leadership theory regarding the importance of shared leadership in knowledge work and teamwork (see Pearce and Sims 2002, Carson et al. 2007). Our findings align with this trend and show that shared and emergent leadership is strong in online communities focused on knowledge exchange. Just as there are several complementary leadership perspectives on leadership in organizations, there is a need to recognize a similar, if not richer, diversity in online settings where individuals generally do not know each other, have ambiguous identities, and are limited in their communications to text-based exchanges. The shared nature of online community leadership is not yet directly recognized in the literature, but is in line with findings about emergent roles, fluid boundaries, and the seeking of position of influence in online communities (Butler et al. 2007, Faraj et al. 2011, Levina and Arrigara 2015).

Third, this study has implications for how researchers study online communities. By asking community participants directly to nominate those they consider leaders, we expand on previous studies of online leadership. For example, Collier and Kraut (2012) and O'Mahony and Ferraro (2007) emphasize formal leadership roles. Huffaker (2010) defines leaders as those who generate the most responses or whose language is most frequently adopted by other participants. Finally, Zhu et al. (2012) focus on leadership behaviors that any community member can perform. Our study offers a direct identification of leaders where peers nominate leaders, and thus it offers a stronger identification approach than those derived from leader activities or network position. We endorse the view that any participant in an online community may demonstrate a leadership behavior (see Zhu et al. 2012), but in considering the most influential participants, we focus on the most active participants. Given that the distribution of online participation follows a scale-free power law (Faraj et al. 2008, Newman 2003), it becomes crucial to carefully select an appropriate subsample for a characteristic of interest, as was done here by comparing participants at similar levels of participation.

Finally, this study's use of NLP provides a set of robust and rich measures to differentiate language use among online community leaders and other participants. With advances in computational linguistics, it is increasingly feasible to collect and analyze large samples of written text. Although future research is needed to determine how our specific findings (e.g., positive, concise posts with familiar language) generalize to other settings, our findings support the application of an NLP prism model utilizing multiple levels of linguistic analysis.

Nonetheless, the exact nature of these patterns, both in terms of communication network structures and linguistic characteristics, may well vary dependent on the context. For example, a closed community handling a crisis situation may have very different patterns of influential communication than a group of hobbyists. Different communities have different values, purpose, and social context that each shape and reinforce the behaviors associated with leadership. Indeed, an open question is to what extent it is possible to theorize about online communities as a unitary phenomenon. Given recent suggestions that online communities differ greatly in terms of regulative behaviors (Kiesler et al. 2011), role taking (Faraj et al. 2011), and social stratification (Levina and Arrigara 2015), it is probable that future research would benefit from approaches that emphasize a richer and more detailed data collection strategy, one that respects the embeddedness and situatedness of the social dynamics in online communities.

Practical Implications

Our results can inform participants and community managers regarding influence and leadership in online communities. First, individual participants can apply the findings of this study. Although occupying a formal leadership role is consistent with being viewed as influential, it is neither a necessary nor a sufficient condition. A participant seeking to become one of the most influential participants of a community should be a highly active participant in many message threads concentrated on closely related topics and should communicate with positive familiar language.

Second, for those who manage or sponsor online communities the findings demonstrate the utility of seeking peer nominations to identify the most influential participants of a community. Whereas the ability to perform a robust linguistic analysis of participant posts is beyond the resources of a typical community manager, participant surveys are quite feasible. Our results demonstrate that, compared to the thousands of participants in the studied communities, even a relatively small number of survey responses can generate a valid list of peer-nominated influential participants. Finally, the findings also demonstrate that when a community manager seeks to reward, incentivize, or for any reason identify the most influential participants, they should not limit themselves to only those already in formal leadership roles.

Limitations and Directions for Future Research

The study design and findings point to multiple avenues for future research. First, more research is warranted to identify additional leadership behaviors and leadership styles associated with online community leadership. These include supporting and leading by example, transaction versus transformative styles, and directive versus empowering communication (O'Donnell et al. 2012, Sims Jr et al. 2009, Yukl 1999). Second, given the wide variety of online communities (Kietzmann et al. 2011), it is quite possible that some traits associated with online community leadership are universal, whereas others are idiosyncratic. Also, future research is needed to identify interactions between formal roles, network configuration or position, and the linguistic characteristics associated with leadership.

Second, although the majority of our propositions are indeed supported, two unexpected findings merit further research. We had expected leaders to use a more sophisticated vocabulary than the average community member because they possibly needed a richer vocabulary to fully answer questions, provide deeper explanations, and delve into discussions about complex knowledge topics. Instead, our measure of vocabulary richness was statistically significant but in the opposite direction than theorized. The results suggest that simpler language, particularly if it is prototypical of community utterance, may be important for effective communication to a wider audience. For the linguistic measures, we found that providing useful resources in the form of Web links did not increase the likelihood of being identified as a leader. We had argued that Web links represent a resource of value to other participants and thus would be perceived as a particularly helpful contribution. The negative finding indicates that Web links may serve as poor proxies for actual resources, may be already known to the recipient, or may be perceived as a throwaway pointer reflecting a lack of engagement in the conversation. It is possible that providing resources is indeed valued, but that Web links serve are a poor proxy of such. In fact, as a measure solely of quantity without regard to quality, posted links could be off topic, self-promotional, or otherwise not of general value.

Finally, this study supports the perspective that leadership in online communities emerges from both the structural and linguistic characteristics of participant communication, but it does not attempt to identify consequences of online community leadership. No doubt, not all online community leaders are equally effective. In addition, the compatibility of individual attributes such as personality, spatial and time separation (Espinosa et al. 2012), and identification with community (Ren et al. 2007) are all likely to impact online community leadership processes. More work is needed to gain a greater understanding of leadership effectiveness in online communities.

Conclusion

This study integrates four perspectives of online leadership—formal leadership roles, peer nominations, network position, and language use—to provide a richer understanding of leadership processes in online settings. Compared to traditional hierarchical organizations, online communities are heavily influenced by emergent leadership processes. To investigate how network structure and language use leads to influence, we analyzed a combination of participant surveys, communication history, and user profiles. Our findings indicate that the participants viewed as leaders not only occupy central, core network positions, but generate distinctive written communication patterns of positive posts using language familiar to other participants. Thus, being an online community leader is associated with both where and how participants post; quantity, position, and quality all matter.

Acknowledgments

The authors thank the senior editor, associate editor, and reviewers for their guidance and valuable feedback during the review process. The authors benefited from insight and assistance from Sri Kudaravalli and Adrian Yeow. The authors also acknowledge the valuable comments provided by participants at McGill University, Temple University, and the University of British Columbia where earlier versions of this work was presented. The third author acknowledges support from the Canada Research Chairs program, and the second and third authors acknowledge support from the Social Science Research Council program.

Appendix. Summary Statistics and Validation

Table A.1 Descriptive Statistics for Analysis Data Set

	Blender Artists		Gearbox Software		Northern Sounds		All	
Sample size	$n = 2{,}101$		$n = 331$		$n = 515$		$n = 2{,}947$	
	Mean	S.d.	Mean	S.d.	Mean	S.d.	Mean	S.d.
Number of posts	131.12	272.09	324.85	348.99	86.85	173.37	145.15	275.73
Centrality	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.01
Coreness	9.51	4.38	19.91	3.84	5.99	1.38	10.06	5.45
Boundary spanning	0.53	0.19	0.52	0.16	0.69	0.15	0.55	0.19
Readability	6.26	2.36	5.47	2.32	6.28	1.87	6.18	2.29
Vocabulary richness	17.86	7.60	10.06	4.96	26.24	11.30	18.45	9.22
Prototypicality	4.64	0.25	4.16	0.62	4.78	0.26	4.61	0.36
External linking	0.22	0.20	0.13	0.12	0.22	0.26	0.21	0.21
Positive sentiment	0.41	0.19	0.24	0.15	0.64	0.30	0.43	0.23

Table A.2 Correlation Table for the Analysis Data Set ($n = 2{,}947$)

	1	2	3	4	5	6	7	8	9	10
1. *Online community leader*										
2. *Formal role of authority*	0.33									
3. *Group membership* (1, 2, 3)	0.07	0.06								
4. *Centrality*	0.39	0.17	0.17							
5. *Coreness*	0.14	0.08	−0.08	0.19						
6. *Boundary spanning*	0.00	0.06	0.30	0.10	0.15					
7. *Readability*	0.01	0.00	−0.02	−0.03	−0.15	−0.04				
8. *Vocabulary richness*	−0.12	−0.06	0.25	−0.22	−0.57	0.08	0.34			
9. *Prototypicality*	−0.02	−0.01	0.03	−0.08	−0.38	0.04	0.20	0.42		
10. *External linking*	0.03	0.03	−0.03	0.00	−0.12	−0.06	0.28	0.19	0.14	
11. *Positive sentiment*	0.00	−0.01	0.28	−0.00	−0.20	0.21	−0.14	0.21	0.19	0.07

Note. Correlations with an absolute value of 0.04 or greater are significant at $p < 0.05$.

Table A.3 Testing for Multicollinearity in the Analysis Data Set Using the Variance Inflation Factor

Measure	VIF	1/VIF
Formal role of authority	1.04	0.96
Centrality	1.10	0.91
Coreness	1.29	0.78
Boundary spanning	1.90	0.53
Readability	1.67	0.60
Vocabulary richness	1.29	0.78
External linking	1.12	0.90
Prototypicality	1.15	0.87
Positive sentiment	1.20	0.83
Mean VIF	1.31	

Table A.4 10-Fold Cross Validation Using the Analysis Data Set

Run	RMSE
1	0.10
2	0.15
3	0.14
4	0.16
5	0.11
6	0.14
7	0.15
8	0.11
9	0.10
10	0.11
Average RMSE	0.13

Table A.5 Descriptive Statistics for the Full Data Set

	Blender Artists		Gearbox Software		Northern Sounds		All	
	Mean	S.d.	Mean	S.d.	Mean	S.d.	Mean	S.d.
Sample size	$n = 10{,}264$		$n = 1{,}644$		$n = 2{,}488$		$n = 14{,}396$	
Number of posts	30.07	133.38	72.30	201.68	20.69	85.81	33.27	137.14
Centrality	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00
Coreness	3.48	3.87	6.91	7.72	2.63	2.08	3.73	4.43
Boundary spanning	0.68	0.28	0.68	0.28	0.75	0.24	0.69	0.27
Readability	6.50	5.14	6.91	25.81	6.62	6.11	6.56	10.07
Vocabulary richness	36.79	30.56	33.96	63.22	50.64	37.86	38.86	37.41
Prototypicality	4.66	0.64	4.22	2.81	4.81	0.49	4.63	1.12
External linking	0.25	1.10	0.17	0.76	0.19	0.44	0.23	0.98
Positive sentiment	0.43	0.45	0.29	0.45	0.57	0.45	0.44	0.46

Table A.6 Correlation Table for the Full Data Set ($n = 14{,}396$)

	1	2	3	4	5	6	7	8	9	10	11
1. *Online community leader*											
2. *Formal role of authority*	0.30										
3. *Group membership* (1, 2, 3)	0.03	0.03									
4. *Number of posts*	0.38	0.14	0.00								
5. *Centrality*	0.40	0.17	0.08	0.75							
6. *Coreness*	0.17	0.10	−0.01	0.55	0.26						
7. *Boundary spanning*	−0.03	0.00	0.10	−0.13	−0.02	−0.18					
8. *Readability*	0.00	0.00	0.01	−0.02	−0.01	−0.03	0.02				
9. *Vocabulary richness*	−0.05	−0.03	0.12	−0.16	−0.08	−0.31	0.17	0.50			
10. *Prototypicality*	0.00	−0.01	0.01	−0.03	−0.01	−0.06	−0.02	0.04	0.16		
11. *External linking*	0.00	0.00	−0.03	−0.01	0.00	−0.02	0.01	0.12	0.15	0.02	
12. *Positive sentiment*	0.00	0.00	0.08	−0.03	0.00	−0.03	0.09	−0.04	0.05	−0.06	0.04

Note. Correlations with absolute an value of 0.02 or greater are significant at $p < 0.05$.

Table A.7 Hierarchical Logistic Regression Model with Dependent Variable *Online Community Leader* Using the Full Data Set

Measure	Model A	Model B	Model C
	Group-level coefficients		
Intragroup correlation	0.106	0.547	0.315
	Participant-level coefficients as odds ratios; standard errors are in parentheses		
Number of posts	1.59*** (0.06)	1.12 (0.06)	1.11 (0.06)
Formal role of authority	72.32*** (35.33)	36.07*** (19.29)	36.85*** (20.29)
Centrality		1.16** (0.05)	1.14* (0.06)
Coreness		4.79*** (1.31)	3.07*** (0.85)
Boundary spanning		0.41** (0.12)	0.41** (0.12)
Readability			1.35 (0.21)
Vocabulary richness			0.05** (0.06)
External linking			1.20 (0.14)
Prototypicality			6.04** (4.00)
Positive sentiment			1.82* (0.53)
	Goodness-of-fit indices		
Log likelihood	−243.1	−203.8	−193.7
χ^2	245.7	173.1	171.1
Akaike information criterion (AIC)	494.2	421.6	411.4
Bayesian information criterion (BIC)	524.5	474.6	502.3
Comparison with previous model ($\Delta\chi^2$)		78.57***	20.27**

Note. $n = 14{,}396$.
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table A.8 Hierarchical Logistic Regression Model with Dependent Variable *Leaders* Using the Analysis Data Set with Moderators and Administrators Removed

Measure	Model A	Model B
	Group-level coefficients	
Intragroup correlation	0.396	0.141
	Participant-level coefficients as odds ratios	
Centrality	1.60*** (0.12)	1.54*** (0.14)
Coreness	4.04*** (1.52)	2.66** (0.98)
Boundary spanning	0.49** (0.11)	0.56** (0.12)
Readability		1.31 (0.19)
Vocabulary richness		0.44* (0.15)
External linking		1.78 (0.18)
Prototypicality		1.75* (0.41)
Positive sentiment		1.43 (0.29)
	Goodness-of-fit indices	
Log likelihood	−184.9	−176.5
χ^2	68.47	76.52
Akaike information criterion (AIC)	379.8	373.0
Bayesian information criterion (BIC)	409.7	432.8
Comparison with previous model ($\Delta\chi^2$)		16.80**

Note. $n = 2{,}925$.
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table A.9 Descriptive Statistics for All Online Community Leaders

	Blender Artists		Gearbox Software		Northern Sounds		All	
	$n = 23$		$n = 21$		$n = 15$		$n = 59$	
	Mean	S.d.	Mean	S.d.	Mean	S.d.	Mean	S.d.
Number of posts	1,088.17	1,392.37	746.91	517.18	608.87	619.28	844.85	980.57
Centrality	0.02	0.04	0.02	0.02	0.03	0.04	0.02	0.03
Coreness	15.35	2.01	20.95	3.83	7.27	0.96	15.29	5.92
Boundary spanning	0.54	0.19	0.47	0.15	0.72	0.13	0.56	0.19
Readability	6.12	1.67	6.25	2.23	6.49	1.87	6.26	1.91
Vocabulary richness	9.34	4.30	9.98	6.41	13.93	6.37	10.73	5.88
Prototypicality	4.62	0.12	4.32	0.42	4.83	0.11	4.56	0.33
External linking	0.23	0.19	0.17	0.10	0.38	0.34	0.24	0.23
Positive sentiment	0.39	0.16	0.23	0.14	0.79	0.29	0.43	0.29

Table A.10 Online Community Participants with Formal Roles of Authority

	Blender Artists		Gearbox Software		Northern Sounds		All	
	$n = 4$		$n = 5$		$n = 3$		$n = 12$	
	Mean	S.d.	Mean	S.d.	Mean	S.d.	Mean	S.d.
Number of posts	967.25	682.11	668.80	322.02	450.33	619.73	713.67	526.92
Centrality	0.016	0.021	0.013	0.006	0.043	0.064	0.022	0.032
Coreness	16.00	0.00	22.20	1.79	7.00	1.00	16.33	6.39
Boundary spanning	0.702	0.13	0.54	0.13	0.78	0.06	0.65	0.15
Readability	5.64	1.64	6.47	2.91	8.87	1.61	6.79	2.44
Vocabulary richness	7.98	2.60	8.30	3.29	16.31	3.41	10.19	4.63
Prototypicality	4.62	0.17	4.24	0.48	4.86	0.028	4.52	0.40
External linking	0.33	0.40	0.15	0.11	0.64	0.41	0.33	0.34
Positive sentiment	0.31	0.24	0.22	0.20	0.74	0.34	0.38	0.31

Table A.11 Comparison of Online Community Leaders With and Without Formal Roles of Authority

Measure	Online community leaders with no formal role $n = 47$; mean (s.d.)	Online community leaders also moderator or administrator $n = 12$; mean (s.d.)	t-test for difference; diff., (t value)
Number of posts	878.3 (1,067.9)	713.67 (526.92)	164.7 (319.2)
Centrality	0.02 (0.03)	0.02 (0.03)	−0.00 (0.01)
Coreness	15.0 (5.8)	16.33 (6.39)	−1.31 (1.92)
Boundary spanning	0.54 (0.19)	0.65 (0.15)	−0.12 (0.06)
Readability	6.13 (1.75)	6.80 (2.45)	−0.67 (0.62)
Vocabulary richness	10.9 (6.2)	10.20 (4.63)	0.67 (1.92)
External linking	0.22 (0.18)	0.33 (0.31)	−0.12 (0.07)
Prototypicality	4.57 (0.32)	4.52 (0.40)	0.05 (0.11)
Positive sentiment	0.45 (0.29)	0.38 (0.32)	0.06 (0.095)

Note. No significant differences were found.

Table A.12 Descriptive Statistics for Online Community Leaders Identified by One Participant

	Blender Artists		Gearbox Software		Northern Sounds		All	
	$n = 21$		$n = 14$		$n = 11$		$n = 46$	
	Mean	S.d.	Mean	S.d.	Mean	S.d.	Mean	S.d.
Number of posts	885.91	931.92	727.36	607.08	352.27	343.87	710.043	751.36
Centrality	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.01
Coreness	15.29	2.10	20.50	4.49	7.00	1.00	14.89	5.75
Readability	6.24	1.63	5.92	1.70	6.33	2.05	6.16	1.72
Boundary spanning	0.52	0.18	0.46	0.18	0.68	0.12	0.54	0.18
Vocabulary richness	9.69	4.30	11.27	7.33	15.53	6.56	11.56	6.23
Prototypicality	4.63	0.11	4.43	0.37	4.83	0.09	4.61	0.26
External linking	0.24	0.19	0.17	0.10	4.83	0.29	0.24	0.20
Positive sentiment	0.38	0.15	0.23	0.14	0.72	0.29	0.41	0.26

Table A.13 Descriptive Statistics for Online Community Leaders Identified by Two or More Participants

	Blender Artists		Gearbox Software		Northern Sounds		All	
	$n = 2$		$n = 7$		$n = 4$		$n = 13$	
	Mean	S.d.	Mean	S.d.	Mean	S.d.	Mean	S.d.
Number of posts	3,212.00	3,924.44	786.00	300.66	1,314.50	700.17	1,321.85	1,488.00
Centrality	0.09	0.12	0.02	0.02	0.09	0.04	0.05	0.06
Coreness	16.00	0.00	21.86	1.95	8.00	0.00	16.69	6.54
Boundary spanning	0.76	0.10	0.48	0.11	0.81	0.08	0.63	0.19
Readability	4.78	2.02	6.90	3.09	6.93	1.37	6.58	2.49
Vocabulary richness	5.68	2.75	7.40	2.98	9.51	3.24	7.78	3.09
Prototypicality	4.51	0.18	4.11	0.45	4.80	0.17	4.38	0.47
External linking	0.13	0.07	0.18	0.11	0.47	0.47	0.26	0.29

Table A.14 Comparison of Online Community Leaders Identified by One Participant to Those Identified by Two or More Participants

Measure	Identified by one participant $n = 46$; mean (s.d.)	Identified by multiple participants $n = 13$; mean (s.d.)	t-test for difference; diff., (t value)
Number of posts	710.0 (751.36)	1,321.9 (1,488.0)	−611.8* (−2.04)
Centrality	0.01 (0.01)	0.05 (0.06)	−0.04*** (−4.61)
Coreness	14.9 (5.75)	16.7 (6.54)	−1.8 (−0.97)
Boundary spanning	0.54 (0.18)	0.63 (0.19)	−0.085 (−1.47)
Readability	6.17 (1.73)	6.58 (2.49)	−0.415 (−0.69)
Vocabulary richness	11.57 (6.23)	7.78 (3.09)	3.78* (2.11)
External linking	0.24 (0.21)	0.26 (0.29)	−0.02 (−0.29)
Prototypicality	4.62 (0.26)	4.38 (0.47)	0.235* (2.35)
Positive sentiment	0.42 (0.26)	0.50 (0.39)	−0.08 (−0.87)

$^{*}p < 0.05$; $^{***}p < 0.001$.

References

Alpaydin E (2004) *Introduction to Machine Learning* (MIT Press, Cambridge, MA).

Aron J (2011) How innovative is Apple's new voice assistant, Siri? *New Scientist* 212(2836):24.

Ashforth BE, Mael F (1989) Social identity theory and the organization. *Acad. Management Rev.* 14(1):20–39.

Balkundi P, Kilduff M (2006) The ties that lead: A social network approach to leadership. *Leadership Quart.* 17(4):419–439.

Barge JK (1994) *Leadership: Communication Skills for Organizations and Groups* (St. Martin's Press, New York).

Barrett DJ (2008) *Leadership Communication* (McGraw-Hill Irwin, Boston).

Barry J (2011) Doing Bayesian data analysis: A tutorial with R and BUGS. *Europe's J. Psych.* 7(4):778–779.

Bass BM, Bass R (2008) *The Bass Handbook of Leadership: Theory, Research, and Managerial Applications* (Free Press, New York).

Benkler Y (2006) *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (Yale University Press, New Haven, CT).

Bird S, Klein E, Loper E (2009) *Natural language processing with Python* (O'Reilly Media, Sebastopol, CA).

Borgatti SP, Everett MG (2000) Models of core/periphery structures. *Soc. Networks* 21(4):375–395.

Borgatti SP, Halgin DS (2011) Analyzing affiliation networks. *The Sage Handbook of Social Network Analysis* (Sage, Thousand Oaks, CA), 417–433.

Bradley MM, Lang PJ (1999) *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings* Technical report C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville.

Burke CS, Stagl KC, Klein C, Goodwin GF, Salas E, Halpin SM (2006) What type of leadership behaviors are functional in teams? A meta-analysis. *Leadership Quart.* 17(3):288–307.

Burt RS (1995) *Structural Holes: The Social Structure of Competition* (Harvard University Press, Cambridge, MA).

Butler BS, Sproull L, Kiesler S, Kraut R (2007) Community effort in online groups: Who does the work and why? Weisband S, ed. *Leadership at a Distance: Interdisciplinary Perspectives* (Lawrence Erlbaum Associates, Mahwah, NJ), 171–193.

Clark A, Fox C, Lappin S (2010) *The Handbook of Computational Linguistics and Natural Language Processing* (Wiley-Blackwell, Chichester, UK).

Cobb NK, Graham AL, Abrams DB (2010) Social network structure of a large online community for smoking cessation. *Amer. J. Public Health* 100(7):1282–1289.

Collier B, Kraut R (2012) Leading the collective: Social capital and the development of leaders in core-periphery organizations. *Proc. Collective Intelligence Conf., Cambridge, MA.*

Cooren F, Kuhn T, Cornelissen JP, Clark T (2011) Communication, organizing and organization: An overview and introduction to the special issue. *Organ. Stud.* 32(9):1149–1170.

Crowston K, Howison J (2005) The social structure of free and open source software development. *First Monday* 10(2).

Dahlander L, Frederiksen L (2012) The core and cosmopolitans: A relational view of innovation in user communities. *Organ. Sci.* 23(4):988–1007.

Donath J (2007) Signals in social supernets. *J. Comput.-Mediated Comm.* 13(1):231–251.

Espinosa JA, Cummings JN, Pickering C (2012) Time separation, coordination, and performance in technical teams. *IEEE Trans. Engrg. Management* 59(1):91–103.

Fairhurst G (2007) *Discursive Leadership: in Conversation with Leadership Psychology* (Sage Publications, Los Angeles).

Faraj S, Johnson SL (2011) Network exchange patterns in online communities. *Organ. Sci.* 22(6):1464–1480.

Faraj S, Jarvenpaa SL, Majchrzak A (2011) Knowledge collaboration in online communities. *Organ. Sci.* 22(5):1224–1239.

Faraj S, Wasko MM, Johnson SL (2008) The structure and processes of electronic knowledge networks. Becerra-Fernandez I, Leidner D, eds. *Advances in Management Information Systems, Knowledge Management: An Evolutionary View of the Field* (M.E. Sharpe, Inc., Armonk, NY), 270–291.

Ferrucci D (2010) Build Watson: An overview of DeepQA for the Jeopardy! challenge. *Proc. 19th Internat. Conf. Parallel Architectures and Compilation Techniques* (ACM, New York), 1–2.

Fleming L, Waguespack DM (2007) Brokerage, boundary spanning, and leadership in open innovation communities. *Organ. Sci.* 18(2):165–180.

Gerstner CR, Day DV (1997) Meta-analytic review of leader–member exchange theory: Correlates and construct issues. *J. Appl. Psych.* 82(6):827–844.

Goldsmith J (2001) Unsupervised learning of the morphology of a natural language. *Comput. Linguist.* 27(2):153–198.

Graen GB, Uhl-Bien M (1995) Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain perspective. *Leadership Quart.* 6(2):219–247.

Gunning R (1969) The fog index after twenty years. *J. Bus. Comm.* 6(2):3–13.

Hackman JR, Walton RE (1986) Leading groups in organizations. Goodman P, ed. *Designing Effective Work Groups* (Jossey-Bass, San Francisco), 72–119.

Hart PE, Duda RO, Stork DG (2001) *Pattern Classification* (John Wiley & Sons, Inc., New York).

Hoch JE, Kozlowski SWJ (2014) Leading virtual teams: Hierarchical leadership, structural supports, and shared team leadership. *J. Appl. Psych.* 99(3):390–403.

Hogg MA (2001) A social identity theory of leadership. *Personality Soc. Psych. Rev.* 5(3):184–200.

Hollingshead AB (2011) Dynamics of leader emergence in online groups. Birchmeier Z, Dietz-Uhler B, Stasser G, eds. *Strategic Uses of Social Technology: An Interactive Perspective of Social Psychology* (Cambridge University Press, Cambridge, UK), 108–126.

Hosmer DW, Lemeshow S (2005) *Applied Logistic Regression* (John Wiley & Sons, Inc., Hoboken, NJ).

Huffaker D (2010) Dimensions of leadership and social influence in online communities. *Human Comm. Res.* 36(4):593–617.

Johnson PD, Dasborough MT (2008) Affective events: Building social network ties and facilitating informal leader emergence. Humphrey RH, ed. *Affect and Emotion: New Directions in Management: Theory and Research* (Information Age Publishing, Charlotte, NC), 175–202.

Jurafsky D, Martin JH (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Prentice Hall, Upper Saddle River, NJ).

Jurafsky D, Martin JH (2008) *Speech and Language Processing*, 2nd ed. (Prentice Hall, Upper Saddle River, NJ).

Kankanhalli A, Tan BCY, Kwok-Kee W (2005) Contributing knowledge to electronic knowledge repositories: An empirical investigation. *MIS Quart.* 29(1):113–143.

Kiesler S, Kraut R, Resnick P, Kittur A (2011) Regulating behavior in online communities. Kraut RE, Resnick P, eds. *Building Successful Online Communities: Evidence-Based Social Design* (MIT Press, Cambridge, MA), 125–178.

Kietzmann JH, Hermkens K, McCarthy IP, Silvestre BS (2011) Social media? Get serious! Understanding the functional building blocks of social media. *Bus. Horizons* 54(3):241–251.

Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS (1975) Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel. Research Branch, Naval Technical Training Command, Millington, TN.

Knoke D, Yang S (2008) *Social Network Analysis* (Sage Publications, Thousand Oaks, CA).

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. Internat. Joint Conf. Artificial Intelligence, Montréal*, 1137–1145.

Kraut RE, Resnick P (2011) *Building Successful Online Communities: Evidence-Based Social Design* (MIT Press, Cambridge, MA).

Lakhani K, von Hippel E (2003) How open source software works: Free user-to-user assistance. *Res. Policy* 32(6):923–943.

Levina N, Arrigara M (2015) Distinction and status production on user-generated content platforms: Using Bourdieu's theory of cultural production to understand social dynamics in online fields. *Inform. Systems Res.* Forthcoming.

Levina N, Vaast E (2005) The emergence of boundary spanning competence in practice: Implications for implementation and use of information systems. *MIS Quart.* 29(2):335–363.

Liu CH (2011) The effects of innovation alliance on network structure and density of cluster. *Expert Systems Appl.* 38(1):299–305.

Long JS, Freese J (2006) *Regression Models for Categorical Dependent Variables Using STATA* (Stata Press, College Station, TX).

Mehra A, Kilduff M, Brass DJ (2001) The social networks of high and low self-monitors: Implications for workplace performance. *Admin. Sci. Quart.* 46(1):121–146.

Mitkov R (2005) *The Oxford Handbook of Computational Linguistics* (Oxford University Press, Oxford, UK).

Morgeson FP, DeRue DS, Karam EP (2010) Leadership in teams: A functional approach to understanding leadership structures and processes. *J. Management* 36(1):5–39.

Newman ME (2003) The structure and function of complex networks. *SIAM Rev.* 45(2):167–256.

Nielsen FÅ (2011) A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.

O'Donnell M, Yukl G, Taber T (2012) Leader behavior and LMX: A constructive replication. *J. Managerial Psych.* 27(2): 143–154.

O'Mahony S, Ferraro F (2007) The emergence of governance in an open source community. *Acad. Management J.* 50(5):1079–1106.

Pang B, Lee L (2008) Opinion mining and sentiment analysis. *Foundations Trends Inform. Retrieval* 2(1–2):1–135.

Pearce CL, Sims HP (2000) Shared leadership: Toward a multi-level theory of leadership. Beyerlein MM, Johnson DA, Beyerlein ST, eds. *Advances in Interdisciplinary Studies of Work Teams*, Vol. 7 (Emerald Group Publishing Limited, Greenwich, CT), 115–139.

Perry ML, Pearce CL, Sims HP Jr (1999) Empowered selling teams: How shared leadership can contribute to selling team outcomes. *J. Personal Selling Sales Management* 19(3):35–51.

Pfeffer J, Cialdini RB (1998) Illusions of influence. Kramer RM, Neale MA, eds. *Power and Influence in Organizations* (Sage Publications, Thousand Oaks, CA), 1–20.

Preece J (2000) *Online Communities: Designing Usability, Supporting Sociability* (John Wiley & Sons, Chichester, UK).

Reagans R, Zuckerman EW (2001) Networks, diversity, and productivity: The social capital of corporate R&D teams. *Organ. Sci.* 12(4):502–517.

Reicher S, Haslam SA, Hopkins N (2005) Social identity and the dynamics of leadership: Leaders and followers as collaborative agents in the transformation of social reality. *Leadership Quart.* 16(4):547–568.

Ren Y, Kraut R, Kiesler S (2007) Applying common identity and bond theory to design of online communities. *Organ. Stud.* 28(3):377–408.

Ren Y, Harper FM, Drenner S, Terveen L, Kiesler S, Riedl J, Kraut RE (2012) Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *MIS Quart.* 36(3):841–864.

Robichaud D, Cooren F (2013) *Organization and Organizing: Materiality, Agency and Discourse* (Routledge, New York).

Royall RM (1986) The effect of sample size on the meaning of significance tests. *Amer. Statistician* 40(4):313–315.

Seidman SB (1983) Network structure and minimum degree. *Soc. Networks* 5(3):269–287.

Shahaf D, Amir E (2007) Towards a theory of AI completeness. *Proc. AAAI Spring Sympos.: Logical Formalizations of Commonsense Reasoning, Palo Alto, CA*, 150–155.

Sims HP Jr, Faraj S, Yun S (2009) When should a leader be directive or empowering? How to develop your own situational theory of leadership. *Bus. Horizons* 52(2):149–158.

Sproull L, Arriaga M (2007) Online communities. Bidgoli H, ed. *The Handbook of Computer Networks: Distributed Networks, Network Planning, Control, Management, and New Trends and Applications*, Vol. 3 (John Wiley & Sons, Hoboken, NJ), 898–914.

Sutanto J, Tan CH, Battistini B, Phang CW (2011) Emergent leadership in virtual collaboration settings: A social network analysis approach. *Long Range Planning* 44(5–6):421–439.

Sy T, Côté S, Saavedra R (2005) The contagious leader: Impact of the leader's mood on the mood of group members, group affective tone, and group processes. *J. Appl. Psych.* 90(2):295–305.

Taylor JR, Van Every EJ (1999) *The Emergent Organization: Communication as Its Site and Surface* (Lawrence Erlbaum Associates, Mahwah, NJ).

Taylor JR, Van Every EJ (2010) *The Situated Organization: Case Studies in the Pragmatics of Communication Research* (Routledge, New York).

von Krogh G, Nonaka I, Rechsteiner L (2012b) Leadership in organizational knowledge creation: A review and framework. *J. Management Stud.* 49(1):240–277.

von Krogh G, Haefliger S, Spaeth S, Wallin MW (2012a) Carrots and rainbows: Motivation and social practice in open source software development. *MIS Quart.* 36(2):649–676.

Walter F, Cole MS, van der Vegt GS, Rubin RS, Bommer WH (2012) Emotion recognition and emergent leadership: Unraveling mediating mechanisms and boundary conditions. *Leadership Quart.* 23(5):977–991.

Wang D, Waldman DA, Zhang Z (2014) A meta-analysis of shared leadership and team effectiveness. *J. Appl. Psych.* 99(2): 181–198.

Warmbrodt J, Sheng H, Hall R (2008) Social network analysis of video bloggers' community. Phuaphanthong T, ed. *Proc. 41st Annual Hawaii Internat. Conf. System Sci.* (IEEE, Los Alamitos, CA).

Wasko MM, Faraj S (2005) Why should I share: Examining social capital and knowledge contribution in electronic networks of practice. *Management Inform. Systems Quart.* 29(1):35–47.

Wasko MML, Teigland R, Faraj S (2009) The provision of online public goods: Examining social structure in an electronic network of practice. *Decision Support Systems* 47(3):254–265.

Weick KE (1969) *The Social Psychology of Organizing* (Addison-Wesley, Reading, MA).

Wickham KR, Walther JB (2007) Perceived behaviors of emergent and assigned leaders in virtual groups. *Internat. J. e-Collaboration* 3(1):1–17.

Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. *IEEE Trans. Evolutionary Comput.* 1(1):67–82.

Yoo Y, Alavi M (2004) Emergent leadership in virtual teams: What do emergent leaders do? *Inform. Organ.* 14(1):27–58.

Yukl G (1999) An evaluation of conceptual weaknesses in transformational and charismatic leadership theories. *Leadership Quart.* 10(2):285–305.

Yukl G (2010) *Leadership in Organizations* (Pearson Education, Upper Saddle River, NJ).

Zhu H, Kraut R, Kittur A (2012) Effectiveness of shared leadership in online communities. *Proc. ACM 2012 Conf. Comput. Supported Cooperative Work* (ACM, New York), 407–416.

Zinsser W (2006) *On Writing Well: The Classic Guide to Writing Nonfiction* (Harper Perennial, New York).